Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 73 (2013) 475 – 478

Procedia
Social and Behavioral Sciences

The 2nd International Conference on Integrated Information

Using Association Rules for Searching Levers of Influence in Census Data

Oleg Chertov*, Marharyta Aleksandrova

National Techical University of Ukraine "Kyiv Polytechnic Institute", 37 Peremohy Prospect, Kyiv 03056,Ukraine

Abstract

Nowadays not only aggregated demographic data but also primary ones become available to a wide range of researches all over the world. It contains a lot of hidden patterns and dependencies which can help answer the question: how to guide human decision-making process in a certain way (for example, whether to have a baby, to move to another place so on). In previous works authors developed a novel Influence search algorithm based on clustering. This algorithm resembles recommendation systems, because as a result it gives a set of rules (recommendations), which can guide a certain social group to a desired state. For instance, if we want to increase birth-rate among young families, we should provide payments for children. As for elder families providing cheap housing loans will be more effective.

In current paper we propose using association rules instead of clustering, as this technique is considered to be among most powerful and popular Data Mining methods. Comparative analysis of the results of both techniques is also given.

Keywords: Data-Mining; association rules induction; census (demographic) data analysis.

1. Introduction

Census data is among the main resources of information about any country. On the basis of census data analysis results the government builds development programs in economic and social spheres, makes new strategic decisions in the management of state affairs. Recently not only aggregated but also primary census data became available to a wide public (for example, IPUMS project [1]). It is obvious that using primary data we can get qualitatively new results and their social importance can be hardly overestimated.

Nevertheless we need to use appropriate analytical tools in order to find hidden patterns in census data. Today mainly statistical methods are used for census data analysis, among them analysis of variance (ANOVA),

* Corresponding author. Tel.: +38-050-333-6710.
E-mail address: chertov@i.ua

doi:10.1016/j.sbspro.2013.02.079

regression analysis, log-linear analysis, nonparametric approaches [2]. These techniques let us verify adequacy of a certain hypothesis. Contrary, Data Mining methods can help find such patterns which weren't even suspected to exist.

In our previous work [3] we introduces a novel Influence search algorithm based on clustering (one of the Data Mining approaches). Proposed algorithm helps identify factors which can stimulate human decision-making process on the basis of comparing characteristics of two contrasting sets. Methodologically it belongs to contrast mining field [4]. Then this algorithm was widened with the fuzzy logic approach [5]. Usage of fuzzy clustering resulted in more precise recommendations given to different social groups. Still clustering is not the only Data Mining approach. That is why in this paper we provide a research of possibility and reasonability of association rules application for census data analysis.

2. Intelligent Analysis of Census Data

2.1. Influence search algorithm

Proposed in [3] Influence search algorithm consists of the following steps:

1. Separate two contrasting groups N_1 and N_2 out of the original dataset. The first group should contain records about those respondents who possess a certain characteristic, and the second one – records about respondents who do not possess it. Additional restrictions can be also imposed on the group definition process (according to the results of the problem domain analysis).
2. Identify those attributes, which can potentially influence the chosen characteristic presence. Mark out attributes for clustering, i.e., attributes which are numerical or can be compared by numbers. Basing on the problem specific define invariant parameters for groups N_1 and N_2.
3. Cluster group N_1, divide it on subgroups.
4. Identify range of values for each invariant parameter corresponding to the subgroups bounds.
5. Using obtained ranges define subgroups prototypes out of the group N_2.
6. Compare characteristics of clustering based subgroups and their prototypes; summarize results.

We applied described algorithm to the California 2000 census 5-percent microfile [6]. The purpose of the experiment was to identify which factors influence human desire to have a baby. As contrasting groups we took N_1 – a set of families with 1 or 2 children aged 0–2 years and N_2 – a set of families without children. Analyzing them we can track family's state conversion from childless to a family with little children. Following parameters were considered as influencing ones (parameters used for clustering are marked with asterisk): home ownership, type of building, number of vehicles available, commercial business on property, spouses' age*, spouses' education*, spouses' ancestry, class of worker for each spouse, husband's total income (in 1999)*. Distribution of families by spouses' ancestry is presented on Figure 1 with ancestry codes revealed in Table 1.

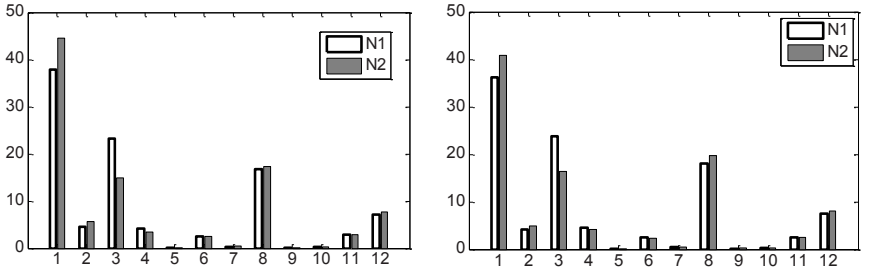

Fig. 1. Distribution of families by husband's (a) and wife's (b) ancestry

Table 1. Ancestry codes

Code	Meaning	Code	Meaning	Code	Meaning	Code	Meaning
1	West Europe	4	Latino	7	other Africa	10	Pasific
2	East Europe	5	Central America Islands	8	other Asia	11	Afro-American
3	Mexico	6	North Africa and South Asia	9	Australia	12	Other American

As we can see most people have West European, Mexican or other (not South) Asian origins, so special attention must be paid to them.

Contrasting groups N_1 and N_2 in [3] were divided on 3 subgroups according to the spouse's age: young families, middle and old. As a result we got following recommendations:

- Providing financial support or cheap housing loans increases birth-rate.
- Providing financial support to the youngest age group representatives with high education level probably won't contribute to their desire having babies, because most highly educated couples decide to have a baby after both spouses are 30 years old.
- If we want couples from the oldest age group to have a baby, we should actively encourage them materially, as lot of families from this subgroup lack of money or own dwelling; besides, if the first child is not born before one of spouses reaches 40 years, the probability of his appearance reduces significantly.

Using fuzzy clustering in [5], we divided original set of respondents on 5 subgroups by husband's age and on 5 more subgroups by wife's age. As a result we got 25 subgroups in total. Distribution of representatives within subgroups and their prototypes is given in Table 2. Among them we took 9 most meaningful ones (Tables 2 in bold) for more detailed investigation and got more accurate recommendations:

- Indeed providing financial support contributes to the birth rate but most people lack own separate dwelling.
- Young age subgroups with low education level should be encouraged by providing cheep house rents (these subgroups are less susceptible to ownership type).
- Young families with high education level usually don't have children, so they shouldn't be considered before spouse become older.
- Special attention should be paid to young couples of Mexican origins, as it is their best reproductive age.
- Families from the middle age subgroups must be actively encouraged with cheap housing loans. Also it should be noted that this period is the most favorable for bearing babies in general, so investment will probably yield the greatest effect here.
- If spouse belong to the elder age subgroups they most likely lack own detached houses. Special attention must be provided to people originated from other (not South) Asia and West Europe.

Table 2. Percentage of families within subgroups of $N_1 - N_2$

Husband's (rows) and wife's (column) age groups	1 (22-25)	2 (26-27)	3 (28-30)	4 (31-32)	5 (33-37)
1 (24-27)	**10.74–9.89**	**4.63–5.51**	2.55–2.78	0.51–0.60	0.39–0.63
2 (28-29)	3.66–3.28	**3.59–4.53**	**6.06–7.04**	1.43–1.43	0.90–1.07
3 (30-31)	1.77–2.13	2.63–2.98	**6.98–7.52**	3.33–3.03	1.86–2.03
4 (32-34)	1.60–1.28	2.13–1.83	**7.18–6.80**	**6.65–6.00**	**8.38–6.95**
5 (35-38)	0.91–0.740.74	1.19–1.09	4.13—4.15	4.41–4.13	**12.27–12.44**

2.2. Mining census data

In this paper we tried to widen proposed Influence search algorithm with association rules mining. First of all we used association rules instead of clustering, that is we tried to mine original subgroups N_1 and N_2 adding to each data record a variable indicating children presence or absence. We used Apriori algorithm implementation in STATISTICA framework and analyzed rules with head "children = yes" and "children = no". Algorithm's input variables were set to: minimum support $supp \geq 1\%$ (support of a certain rule doesn't have to be large because we are looking for meaningful recommendations concerning little subgroups); minimum confidence $conf \geq 70\%$ (rules must have strong influence on a desire to have a baby); minimum correlation $cor \geq 10\%$. We got 3 rules with head "children = yes" (positive rules) and 781 rules with head "children = no" (negative rules).

Obtained rules analysis didn't result in more precise recommendations comparing with [5]. This showed that using even such intelligent technique as association rules mining in its pure form can be useless. So we decided to use this method not as an alternative to clustering but as its supplement. On the step 6 of the Influence search algorithm we used association rules to specify result of the paper [5]. Beforehand all 9 subgroups were divided on "subsubgroups" according to the spouses' ancestry. We singled out pure families (husband and wife have the same meaning of the ancestry code) with West European, Mexican and not South Asian origins. Minimum support was set to 5% as each analyzed subgroup became considerably smaller.

Some rules provided new valuable knowledge, for example a pair of rules for Mexican families of subgroup 3-3:

- IF <wife work class=employee of private for profit company> AND <detached house=yes> THEN <children=yes> ($supp \geq 21.5\%$, $conf \geq 72.7\%$)
- IF <wife work class=employee of private for profit company> AND <detached house=no> THEN <children=no> ($supp \geq 19.5\%$, $conf \geq 72.5\%$)

shows that if we'll provide a detached house to Mexican families from subgroup 3-3 with wives being employee of a private for profit company, we can be sure with a high confidence (72.7%) that they will born a baby.

3. Conclusions

This paper is a continuation of our work series [3, 5] concerning influence levers searching in census data. Previously we developed a clustering based algorithm, which was now improved with association rules mining. This work also showed that using even powerful techniques in their pure form can be useless, but their combinations can result in new knowledge.

Developed Influence search algorithm provides a set of recommendations concerning human behavior influence, so it seems promising to adapt it to classical tasks of recommendation systems field.

References

[1] Minnesota Population Center, University of Minnesota. Integrated Public Use Microdata Series International. https://international.ipums.org/international/

[2] U.S. Census Bureau. Statistical Quality Standard E1: Analyzing Data. http://www.census.gov/quality/standards/standarde1.html

[3] Chertov, O., & Aleksandrova, M. (2011). Clustering with Prototype Extraction for Census Data Analysis. Proceedings of the World Conference on Soft Computing, WConSC-2011, San Francisco, CA, USA, http://arxiv.org/abs/1106.5122

[4] Dong, G., & Bailey, J. (2011). Overview of Contrast Data Mining as a Field and Preview of an Upcoming Book. IEEE 11th International Conference on Data Mining Workshops, Vancouver, Canada.

[5] Chertov, O., & Aleksandrova, M. (2012). Fuzzy Clustering with Prototype Extraction for Census Data Analysis, in book: Soft Computing: State of the Art Theory and Novel Applications, Springer-Verlag, 2012. To be published.

[6] U.S. Census 2000. 5-Percent Public Use Microdata Sample Files. http://www.census.gov/Press-Release/www/2003/PUMS5.html